Before the
FEDERAL COMMUNICATIONS COMMISSION
Washington D.C. 20554

In the Matter of	)	
	)	
A National Broadband Plan	)	GN Docket No. 09-51
For Our Future	)	

REPLY COMMENTS OF ONLIVE, INC.

Quality of Service Metrics and the National Broadband Plan

I. Introduction to OnLive, Inc. as an Exemplary Future Internet Application

OnLive, Inc. ("OnLive") (www.onlive.com) is a Silicon Valley startup delivering real-time interactive games, graphics-intensive applications, and media: delivered from "cloud-based" (datacenter-based) computers to users over the public internet. This notably includes the OnLive® Game Service, a "cloud computing" video game and application on-demand platform, launching later this year. Using a consumer's internet connection, games and other applications which demand high-end graphics and perceptually instantaneous-response interactivity can today be delivered directly to TVs via OnLive's inexpensive MicroConsole™, or directly to a PC or Mac using a small web browser plug-in.

OnLive's technology is an illustration of advanced services that will be soon running on US broadband internet connections going forward, particularly for consumers and small businesses, and as such, OnLive can provide a helpful point of reference as the FCC shapes its National Broadband Plan. In review of previously submitted comments to the FCC, OnLive is concerned that while some of the recommendations will adequately support prior internet applications, they will not support some future applications such as OnLive.

Described below is a summary of OnLive's service offering. (More details on OnLive's technology are provided at the end in Appendix A.).

OnLive is a new type of "cloud computing" system: an "interactive cloud computing system", one in which the "cloud computing" (i.e. computing on a server in the internet) is indistinguishable from what computing experience would be if the application were running entirely on a local computer.

Generally speaking, a cloud computing system is one where some part of the computing happens remotely through the internet (aka "the cloud"). For example, using your browser to do word processing using an internet-based word processing service is an example of cloud computing: The word processor application is running on a server in a data center connected to the internet, perhaps thousands of miles away, rather than on your local computer.

Steve Perlman
OnLive, Inc.

The OnLive cloud computing experience is perceptually indistinguishable from a local computing experience: When you perform an action, you experience the result of that action immediately, just as if the application was running locally. But OnLive is not limited to providing a snappy experience for just simple productivity applications. OnLive was designed to meet the requirements of the most demanding of applications: fast-action video games. Not only do fast-action video games require perceptually instantaneous response times, but they are very demanding in terms of the scene complexity, erratic motion and unpredictability of real-time visual imagery. Indeed, video games create more sudden and topsy-turvy motions than even high-action movies.

Described below are numerous benefits to the public interest associated with the new type of interactive cloud computing services which the OnLive platform exemplifies. In order to take advantage of these benefits, consumers certainly must have ubiquitous broadband internet access, which the FCC has been supporting for years. However, the demands of new kinds of services unsurprisingly place a spotlight on network performance aspects which have previously received essentially no attention. The most important of these aspects is the need for a high and consistent Quality of Service ("QoS") associated with an internet connection.

II. FCC Has Opportunity to Shape the Deployment of Broadband

The FCC's focus to date has been to foster the build-outs of broadband networks. OnLive certainly supports accelerated deployment of broadband serving the under- and un-served. Broadband vendors have been vigorous in trying to serve extant demand: demand in prior years has been primarily focused on getting any broadband access at all, and next increasing data rates to meet increasingly bandwidth-demanding applications, such as video and large downloads.

But in shaping the future of broadband deployment, the FCC must not limit its analysis to prior generation internet applications in a way that would inhibit innovation. We can already see tomorrow's demands struggling with heretofore unheralded network limitations. Thus the FCC should remain focused on the most important element at this juncture—establishing a fundamental approach that is not nearsighted as it accomplishes availability goals, but that also takes into account future developments and needs that service applications will place upon a broadband network. At the same time, the FCC must set its sights high to make way for future expansion and development of technologies. Flexibility and high standards will preserve opportunities to realize the future potential for economic expansion based on new technologies. A truly nationwide high-speed and high-quality broadband build out will enable many advances and benefits to consumers that can't even be imagined at this time.

While we cannot see perfectly into the distant future, we certainly can and must recognize the significant developments and innovations which we do have at hand or can see immediately before us. To ignore them would be to relegate our plans to obsolescence before deployment. And, as the OnLive service shows, network QoS aspects—such as latency, packet loss, and jitter—have been promoted from yesterday's technical trivia to today's critical criteria for US consumers, schools and businesses.

Thus, in addition to insisting upon widely available, competitive internet access to all Americans, The National Broadband Plan also needs to define, describe and report the quantitative QoS items including latency, packet loss, jitter, contention levels, and availability.

III. Public Benefits

There are two kinds of public benefits of ensuring the necessary qualities of broadband development: immediate benefits at hand and vastly more interesting benefits we cannot yet know with precision.

A. Benefits At Hand

Cloud computing technology, both in configurations such as that of the OnLive service and in configurations offered by other data center-based service offerings offers a multitude of the public benefits that should be supported by national broadband infrastructure. For instance, some of the benefits realized by using cloud computing include:

1. Access to high-end applications to all connected users without the need of high-cost, high-power and high-maintenance computers in the home or workplace. This will reduce costs to homes and businesses and reduce residential and business energy consumption.

2. Removing the threat of piracy of software applications and interactive content since the software resides in the cloud (server center) and only the real-time display information generated by the application is provided "over the wire" to end users. This will serve to promote the development and investment in software, a core area of leadership by US companies and a major driver of US economic growth.

3. All users have access to the latest state of the art software instantly and universally (no upgrades or patches needed).

4. Secure storage of data and information in a safe, secure and efficient data center with access from anywhere there is broadband internet connectivity.

5. Improved energy efficiency as users will have simpler, cheaper and less costly user machines with computing, storage and processing power concentrated in a data center where a multitude of user can access and use the same piece of high-end computing hardware and systems. Financial and environmental costs of computers are shared among many users (e.g. with a typical computer are shared amongst multiple users at different times or at the same time through virtualization) instead of incurred by each individual user and are thereby dramatically reduced overall.

6. Users can dispense with complex computer purchases, concomitant computer upgrades, software installs, and related IT work and maintenance.

B. Uncertain, Yet Likely Benefits

We know that when disruptive new technologies become available through widely-available infrastructure, the largest benefits are almost never imagined up front. For example in 1994, when Netscape arrived, even a very imaginative person would have never have imagined today's internet, let alone the value of today's internet. It would have been easy to underestimate the significance of any action which failed to support the development of the internet infrastructure.

Similarly, we cannot now imagine where a truly interactive cloud infrastructure will take us. We see significantly improved ease, economics, ubiquity, IP protections, and manageability of systems, but we know these form only the tip of the iceberg.

However, we can confidently predict some of those applications:

1. Providing schools with up-to-date, state-of-the-art computing resources for students. No longer will schools be faced with the constant upgrade and maintenance of computers and software. A monitor (with built-in thin client), keyboard and mouse (or future pointing devices) will be all that is needed, and a school will always have high-performance computing with current software.

2. Providing up-to-state, secure computing resources for low-income households.

3. Creating a new vehicle for creative development and distribution of computer-generated movies, animations, educational documentaries and games, that normally would be too expensive to develop for independent artists, filmmakers, teachers and game developers.

4. Providing secure computing for both military and civilian applications: since sensitive data is only stored remotely in the data centers, if a laptop, cell phone or other device is stolen, the data remains secure.

5. Reducing the cost of the massive computing resources required for the development of new drugs, materials and technologies.

6. Supporting interactive telemedicine to remote locations lacking specialists, potentially even enabling remotely-controlled robotic surgical techniques.

Also, we do know that in order for this new kind of remote computing capability to achieve widespread reality, and for development of those kinds of applications to be possible, broadband infrastructure must be able to support the demands of cloud computing applications such as those supported by OnLive.

IV. Interactive Cloud Computing

The vast majority of current services, applications and media available on the internet use existing infrastructure and its inherent limitations exceedingly well. These applications generally are those that are largely unidirectional and with loose response deadlines; they download software, content and media objects based on limited amount of user interaction. Other applications from the web download executable programs which are then run in a user's local machine environment, using the internet only for a limited exchange of data and commands. This methodology requires an end-user machine to have the full extent of computing power (*e.g.*, processor, memory, storage and graphics) as well as entire programs to be downloaded into the local user environment.

As true interactive cloud computing becomes available, expensive hardware, software, data, and complex processes can stay in the data center. This reduces the need, cost, complexity and energy consumption of end user computers. Further, when central systems are shared by many users, any negative impacts associated with those systems are divided amongst many users: not only are economic costs shared, but environmental costs like energy use and environmentally unfriendly materials used in the production of computer systems are shared.

As with any technological evolution, there are increased requirements on certain system components. One of those systems called on for increased performance is the internet connection between the end user and the data center. In the past, "performance" pretty much meant "bandwidth," and things that affect bandwidth. Today, performance aspects with little previous attention are noticeably key performance items. This is largely because the interactive cloud-based applications process and provide (render) the video and display images in the data center, those images need to be processed, compressed and conditioned to be transmitted to the end user as quickly as possible, and the user is providing real-time feedback (via mouse, controllers and keystrokes) based on those real-time-provided images.

A. Performance Metrics relevant to both prior and future internet applications:

    1. Bandwidth (*i.e.*, data throughput)

    2. Availability

    3. Price

B. Performance Metrics which are particularly relevant for future internet applications, such as interactive cloud computing, include:

    1. Latency: the delay when packets transverse the network, measured using Round Trip Time (RTT). Packets can be held up in long queues, or delayed from taking a less direct route to avoid congestion. Packets can also be re-ordered between the transmission and reception point. Given the nature of most existing internet applications, latency is rarely noticed by users and then only when latency is extremely severe (seconds). Now, users will be noticing and complaining about latencies measured in milliseconds because of the

accumulation of latency as messages route through the internet, and the immediate-response nature of interactive cloud computing.

2. Jitter: random variations in latency. Prior-technology internet applications used buffering (which increased latency) to absorb and obscure jitter. As a result, users have not noticed or cared about jitter, and the common preconception is that jitter is a technical detail that has no impact on user experience or the feasibility of provisioning internet applications. With interactive cloud computing, excessive jitter can have a significant impact on user experience and perceived performance, ultimately limiting the range of applications.

3. Packet Loss: data packets lost in transmission. In the past, almost all internet traffic was controlled by TCP (Transmission Control Protocol), which hides packet losses by asking for retransmissions without the user's knowledge. Small packet losses come with small increases in latency and reductions in bandwidth, essentially invisible to users. Large packet losses (several percent and up) felt like a "slow network" not a "broken network." With interactive cloud computing the additional round-trip latency delay incurred by requesting a resend of a lost packet potentially introduces a significant and noticeable lag.

4. Contention: multiple users competing for the same bandwidth on an ISP's (Internet Service Provider) network in excess of the network's capacity, without a fair and consistent means to share the available throughput. As applications and use of internet infrastructure continue to grow, old assumptions about the rarity or improbability of contention are being overturned. Contention leads to exacerbation in all three areas: latency, jitter and packet loss, mentioned above.

V. A National Broadband Plan Should Include a Range of Metrics

Many of the submitted comments in this proceeding focused on provisioned bandwidth speed. That is just one of several metrics one must consider in enabling high quality broadband services. Quality of Service metrics (QoS) should be considered as part of a service definition by an internet provider to an end consumer.

Several countries, including Japan, Singapore and the Republic of Korea, have established broadband QoS standards and these have led to greater broadband availability, greater consumer acceptance and a wider availability of applications and services being available to end consumers. Such data is typically already collected by ISPs for the internal maintenance, load balancing and planned built-out of their infrastructure. Also, such information is often provided to commercial customers. So, the publishing of this largely available information should not incur significant cost to ISPs, but it will open the door for a wide range of new applications.

Further, it is seldom feasible for users to measure the QoS of consumer and small business connections, since making such measurements accurately would generally require users to have a

commercial server installation within the internet with very precisely-understood connections to their service providers. Without detailed QoS information, consumers and small businesses will lack the information they need to assess the quality of their internet service in accordance with their needs, particularly as internet applications evolve. For example, a less expensive broadband service offering with poor QoS may be entirely adequate for a user with light web surfing and email needs, but a more expensive broadband service offering may be essential for a consumer running a home business or a school relying upon remote interactive cloud computing. But, if the consumer has no knowledge of the QoS of a broadband services offering, no such assessment can be made. Indeed, current advertised information about broadband service offerings may be confusing or misleading. For example, there are often significant discrepancies between advertised peak bandwidth and typical sustained bandwidth. The FCC currently has no QoS reporting requirements for ISPs. We recommend that the FCC establish the requirement that QoS standards for internet providers be provided to end consumers and reported on annual basis as part of the National Broadband Plan. QoS Reporting Requirements should include each of:

- Standard QoS parameters, such as those discussed previously, including accurate characterization of bandwidth
- Best and worst cases
- Statistics, illustrating the ongoing performance of the broadband service offering

for each of

1. Bandwidth: throughput in bits per second
2. Latency: time to traverse the ISP's network, both upstream and downstream, between hops, and in the "last mile" to the end-user
3. Packet Loss: percentage of packets lost before retransmission attempts
4. Jitter: packet arrival time deviation
5. Availability: percent of time services meet QoS standards
6. Contention: ratio of the collective potential demand at rated capacity to the actual provided capacity. As an example, consider 10 households each provided with 10Mbps bandwidth served by a common line capable of carrying 25 Mbps. If all 10 households maxed their lines simultaneously, the resulting 100Mbps demand would be 4 times the capacity of the line, yielding a contention ratio of 4.0.

Finally, the FCC should consider whether Form 477 should include a report of ISPs QoS performance.

Respectfully submitted,

Steve Perlman
Founder, CEO and CTO
OnLive, Inc.
181 Lytton Ave.
Palo Alto CA 94301
Tel. (650) 543-5508
webmail-fcc@onlive.com

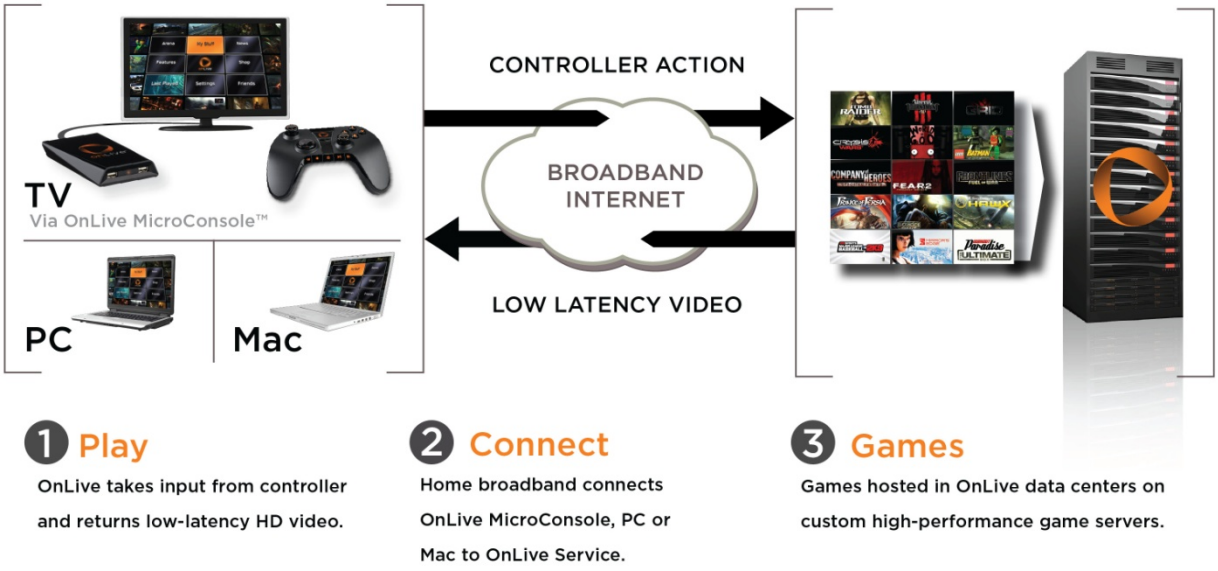
Appendix A: Detailed Description of OnLive Service and Technology

The service and applications that OnLive are implementing are illustrative of future internet applications that will be developed and available to consumers. OnLive's service and technology is described below for illustrative purposes to exemplify the critical nature of the above described internet performance specifications and issues.

.

A. OnLive at a high level

OnLive system architecture is as follows:



When the user performs an action on a computer or TV connected to OnLive (e.g. presses a button on controller or moves a mouse) that action is sent up through the internet to an OnLive data center and routed to a server that is running the game the user is playing (or the application the user is using—since the interactive demands for video games are generally higher, remote video game operations will be primarily described in the following paragraphs, but these discussions are entirely applicable to remote application operations). The game computes the next video frame based on that action, then a proprietary chip compresses the video from the server very quickly, and the user's PC, Mac or OnLive MicroConsole™ decompresses the video and displays the new frame of video on the user's computer display or TV set. The entire round trip, from the point the button is pressed to the point the display or TV is updated is so fast that, perceptually, it appears that the screen is updated instantly and that the game is actually running locally.

The key challenge in any cloud system is to minimize and mitigate the issue of perceived latency to the end user.

Steve Perlman
OnLive, Inc.

B. Latency perception

Every interactive computer system that is used, whether it is a game console, a PC, a Mac, a cell phone, or a cable TV set-top box, introduces a certain amount of latency (i.e. lag) from the point you perform an action and you see the result of that action on the screen. Sometimes the lag is very noticeable (e.g. on some TV set-top boxes it takes over a second to move a selection box in a program guide). Sometimes it isn't noticeable (e.g. if you have a well-designed game running on fast hardware, and pressing the fire button results in what appears to an instantaneous display on your screen of the your gun firing).

But, it's important to note that, even when your brain perceives game response to be "instantaneous", there is always a certain amount of latency from the point you perform an action and your display shows the result of that action. There are several reasons for this. To start with, when you press a button, it takes a certain amount of time for that button press to be transmitted to the computer or game console (it may be less than a millisecond (ms) with a wired controller or as much as 10-20 ms when some wireless controllers are used, or if several are in use at once). Next, the game needs time to process the button press. Games typically run between 30 and 60 frames per second (fps), so that means they only generate a new frame every $1/30^{th}$ to $1/60^{th}$ of a second (33ms to 17ms). (Further, when games are generating complex scenes, sometimes they take longer.) So, even if the game responds right away to a button action, it may not generate a frame for 17-33ms or more that reflects the result of the action. And, then finally, there is a certain amount of time from the point the game completes generating the frame until the frame appears on your display. Depending on the game, the graphics hardware, and the particular monitor you are using, there may be almost no delay, to several frame times of delay. And, if your game is an online game, there typically will be some delay to send a message reflecting your action through the internet to other game players, and the game may (or may not) delay the action occurring in your game so as to match your screen action to that of screen action of players who are playing the game remotely. So, in summary, even when you are running a game on a local machine there is always latency. The question is simply how much latency.

So, while there certainly are more subtleties to the perception of latency, as a general rule of thumb, if a player sees a fast-action game respond within 80ms of an action, not only will the player perceive the game as responding instantaneously, but the player's performance will be just as good as if the latency was shorter.

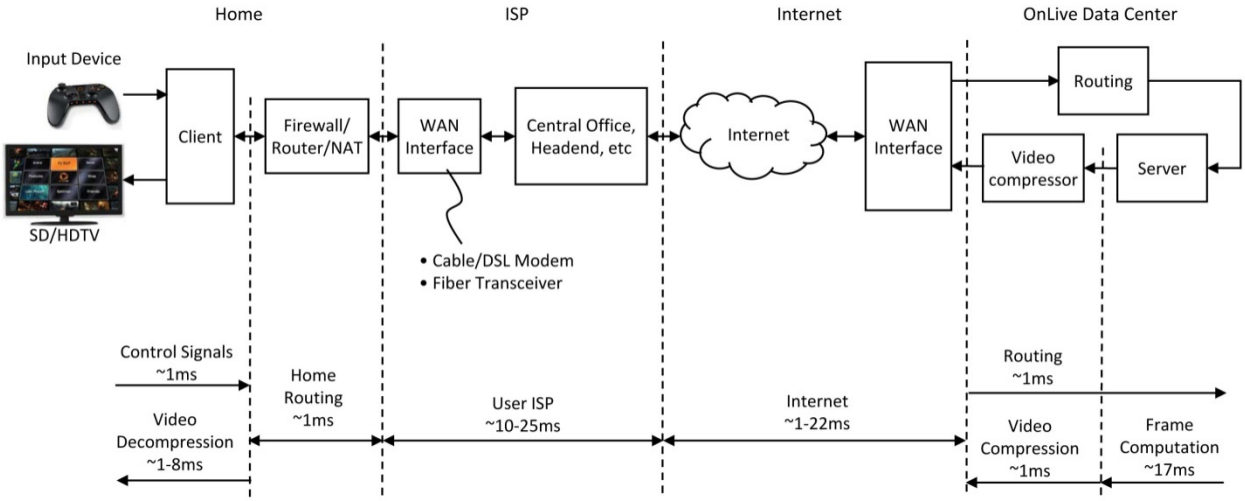
And, as a result, 80ms is the "latency budget" needed to meet for the OnLive system to be practical. That is to say, OnLive has up to 80ms to: send a controller action from the player's home through the internet to an OnLive data center, route the message to the OnLive server running the game, have the game calculate the next frame and output the video, compress the video, route the compressed video out of the data center, send the compressed video to the player's home through the internet, decompress the video on the players computer and output the video to the player's display. And, of course, OnLive has to do this at rate of 60fps with HDTV resolution video over a consumer internet connection, running through consumer internet gear in the home.

Over Cable and DSL connections, OnLive is able to achieve this if the user's home is within about 1000 miles of the OnLive data center. So, through OnLive, a user who is 1000

miles away from a data center can play a video game running on a server in the data center with the perception (and the game play score) as if the game is running locally.

C. OnLive's latency calculations

The simplified diagram below shows the latencies encountered after a user's action in the home makes it way to an OnLive data center, which then generates a new frame of the video game and sends it back to the user's home for display. Single-headed arrows show latencies measured in a single direction. Double-headed arrows show latencies measured roundtrip.



The latency numbers shown here are numbers that OnLive has seen in practice, given the way the OnLive system was architected and optimized, and reflect what has been measured after using OnLive in various locations over the years. If you add up all of the worst-case numbers, it shows the latency can be as high as 80ms. That said, it is highly unlikely that every segment will be worst case so the total latency will likely be much less (and indeed, that is what we see in practice).

D. ISP latency

Potentially, the largest source of latency is the "last mile" latency through the user's Internet Service Provider (ISP). This latency can be mitigated (or exacerbated) by the design and implementation of an ISP's network. Typical wired consumer networks in the US incur 10-25ms of latency in the last mile, based on OnLive's measurements. Wireless cellular networks typically incur much higher last mile latency, potentially over 150-200ms, although certain planned 4G network technologies are expected to decrease latency.

Within the internet, assuming a relatively direct route can be obtained, latency is largely proportional to distance, and the roughly 22ms worst case round-trip latency is based on about 1000 miles of distance (taking into account the speed of light through fiber, plus the typical delays OnLive has seen due to switching and routing through the internet.

Consequently, OnLive will be locating its data centers such that the distance to most of the US population is less than 1000 miles

Steve Perlman
OnLive, Inc.

The compressed video, along with other data required by the OnLive client to keep it tightly synchronized with the OnLive service, is then sent through the internet back to the user's home. Notably, the data generated by the video compressor is carefully managed to not exceed the data rate of the user's internet connection because if it did, that might result in queuing of packets (incurring latency) or dropped packets. Since the user's home data rate is constantly changing, the OnLive service is constantly monitoring the available data rate, and constantly adapting the video compression (and if necessary, dropping the video resolution) to stay below the available data rate.

One common misconception about home broadband connections is that the latency is directly tied to data rate (i.e. the effective connection speed) and/or data throughput (i.e. the data rate available to a particular user). Latency is actually largely independent of data rate, so long as the data throughput demands are less than the capacity of the broadband connection.

E. OnLive video decompression latency

Once the compressed video data and other data is received by the OnLive client (i.e. the OnLive application running as a plug-in or standalone in your PC or Mac, or the OnLive MicroConsole attached to your TV), then it is decompressed. The time needed for decompression depends on the performance of your PC or Mac (CPU and frame buffer bandwidth…no GPU is needed), and may vary from about 1 to 8ms. If your computer's CPU and/or memory bus is tied up doing another processing-intensive task or if you have an extremely low performance computer, OnLive may find it is unable to decompress video at full screen resolution. If so, then it will scale down the video window accordingly. But, we have found most computers made in the last few years work fine up to their screen resolutions so long as they are not tied down running some other intense application at the same time. In any case, even if you are in a processing-constrained situation, OnLive will select a video frame size which will maintain low latency.

F. OnLive round-trip latency

As mentioned before, while there is a certain amount of latency variability in each leg of the journey, it is rare that a given user will end up in a worst-case scenario with each leg. Consequently, what we typically see in practice are latencies on the order of 40 to 60ms. Sometimes we see latencies that are higher and sometimes we see latencies that are shorter. And, we expect latencies to continue to decline as "last mile" infrastructure is upgraded, both for wired and wireless networks.